INFERRING THE COMMANDS OF AN ARTICULATORY MODEL FROM ACOUSTICAL SPECIFICATIONS OF STOP/VOWEL SEQUENCES

R. Laboissière and A. Galvan

Institut de la Communication Parlée
URA CNRS 368 / INPG / Univ. Stendhal
46, av. Félix Viallet 38031 Grenoble CEDEX 1 France

E-mail: `rafael@icp.grenet.fr`

ABSTRACT

We present a part of our efforts towards an articulatory speech synthesizer capable of learn to produce articulatory gestures from acoustical description of the tasks. We concentrated on the problem of characterizing stop consonants in the formant space. We model the stop consonants targets as probabilistic models, which has advantages for both a quantitative assessment of the principle and for application to a model of speech motor control.

INTRODUCTION

A model was developed for obtaining the commands of an articulatory model of the vocal tract from acoustical targets (Laboissière 1993). Using this model, acceptable vowel–vowel transitions are obtained and typical phenomena related to coarticulation and compensation for perturbation can be replicated. These satisfactory results rely on the fact that acoustical targets (in our case the three or four lowest formants) are well defined for vowels. Problems arise when trying to use the system to infer articulatory commands for stop/vowel transitions.

Attempts to find invariant acoustical cues for stop consonants are abundant in the literature (Stevens and Blumstein 1978; Kewley-Port 1983; Sussman et al. 1991). Although the invariance at the acoustical level is still a matter of debate, we are pursuing this paradigm in order to test its validity in the context of a model of motor control for speech production.

In this paper we will describe the preliminary efforts towards our approach to vowel-consonant-vowel articulatory synthesis, and is organized as follows: we present first the principles of our inversion model; second, the technique for obtaining targets for consonants in the acoustical space will be presented as well as an preliminary assessment of the principle.

THE INVERSION MODEL

The schematic of the model we are using to invert from acoustical (distal) desired outcomes into articulatory (proximal) commands is shown in Fig. 1. This scheme is reminiscent of classical techniques in Control Theory, namely feedback control with learning of a feedforward controller.

This control model drives an articulatory model of the human vocal tract (Maeda 1988, $F(u)$ in Fig. 1), implemented as a computer program. The articulatory model was generated from cineradiographic data from a speaker uttering ten phonetically-equilibrated French sentences. From a sort principal component analysis of the mid-sagittal tongue contour it was possible to derive seven articulatory commands like jaw/tongue position, lips aperture/protrusion and larynx height (these commands compose the vector u). At the output, after computing the area function of the resulting configuration of the vocal tract, we extract the first four formants (y in Fig. 1).

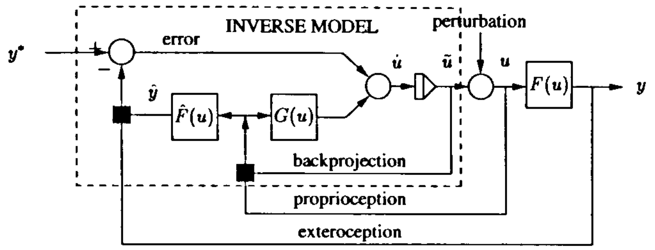

Figure 1: Architecture of articulatory controller. The system to be controlled (the plant) is indicated by $F(u)$, the articulatory inputs are u and the perceptual outputs (formants) y. The *inverse model*, capable of inferring the articulatory inputs from the desired outputs y^* contains a forward model $\hat{F}(u)$ that gives estimations $\hat{y}$ of the plant outputs from the articulatory inputs, obtained either by proprioceptive feedback or through "backprojection."

As the number of inputs is greater than the number of outputs—i.e. the system has degrees of freedom in excess—there is no unique inverse transformation from the desired formants into articulatory commands. The proposed architecture solves this problem in two steps. First, a *forward model* of the articulatory model is learned ($\hat{F}(u)$) in order to mimic the articulatory model [see Jordan and Rumelhart (1992) for a thorough discussion on forward modelling]. More precisely, $\hat{F}(u)$ is an analytical approximation of the mapping $F(u)$. To find this regression model, we are using a mixture of linear experts trained by the Expectation-Maximisation (EM) algorithm, a technique introduce by (Jordan and Jacobs 1994). Essentially, the forward model implements a piecewise linear function.

The main interest of having the piecewise linear approximation (or any simple regression) resides in obtaining a simple expression for the controller $G(u)$. Indeed, $G(u)$ implements a piecewise constant matrix of transformation between the vector of error in the acoustical space (derived from both y^* and $\hat{y}$ or y) and the changes in the articulatory commands. As we use the pseudo inverse of the JAcobian, we ensure that minimal changes in the articulatory variables will be produced for a given acoustical error vector (Klein and Huang 1983). This means that our model can produce smooth commands without any need for planning. Another interesting feature of this architecture is that once the forward model has been learned, the combination of $\hat{F}(u)$ and $G(u)$ can act as a feedforward *inverse model.*

ACOUSTICAL TARGETS FOR STOP CONSONANTS

Let us turn now on how the controller shown in Fig. 1 actually works. In order to obtain articulatory movements, we have to present targets y^* at the input. For vowel production, this targets could be simply formant values (F_1 to F_4) and the error would be some distance between y^* and either y or $\hat{y}$. For the stop consonants there is no target in the formant space due to the occlusion of the vocal tract. The cues that convey information on the stop consonant identity are numerous, ranging from formant transitions to burst spectra [see Kewley-Port (1983) for a review].

In the present work, instead of concentrating on a dynamical description of stop–vowel production, we are asking a more fundamental question: is it possible to identify place of constriction from a kind of "intended formant configuration" that would be produced by the vocal tract just at the moment of occlusion release? Of course, this "formants" would not exist physically in the speech signal, but could be considered as intentional target for stops.

This should relate to the locus equations (Sussman et al. 1991), but we are interested in a more general result, in which stop consonants targets could be associated with large regions in the formant space. We did a thorough exploration of the articulatory model, and were able to obtain several articulatory configurations that give the same place of constriction for the tongue. By computing the formant values for those configurations assuming a small aperture at the place of constriction, it is possible to obtain sets of points in the formant space like those shown in Fig. 2 (only F_2, F_3, and F_4 are shown, because F_1 is systematically close to 200 Hz for all configurations). The big variability observed is due to the free articulators, like lips and larynx, as well as to compensations between jaw and tongue positions.

The case shown in Fig. 2 is quite instructive. The clouds correspond to the same place of constriction (about 1.5 cm behind the upper incisors) but produced with different parts of the tongue: either the tongue dorsum or the tongue tip in a retroflex articulation. We see that the clouds are quite separable, but a more quantitative and systematic assessment of this assertion is called for. In order to do that, we model the cloud of points in the F_2–F_4 space as a probabilistic model, namely as a mixture of Gaussians. Given a vector y in that space and a model $\mathcal{M}_j$ related to a given position of constriction and mode of articulation (tongue tip or tongue dorsum), the probability of having y associated to $\mathcal{M}_j$ is given by

$$P(y|\mathcal{M}_j) = \sum_i g_{ji}|C_{ji}|^{-1/2}e^{-(y-y_{ji})^T C_{ji}^{-1}(y-y_{ji})}, \tag{1}$$

where g_{ij} are the a-priori probabilities, y_{ij} the mean vectors and C_{ij} the covariance matrices. For each of the possible locations of constriction of the articulatory model (from the alveolar to velar regions) spaced by 0.5 cm we found the best mixture of Gaussians using the EM algorithm. We observed that 4 Gaussians were in general sufficient for describing each cloud.

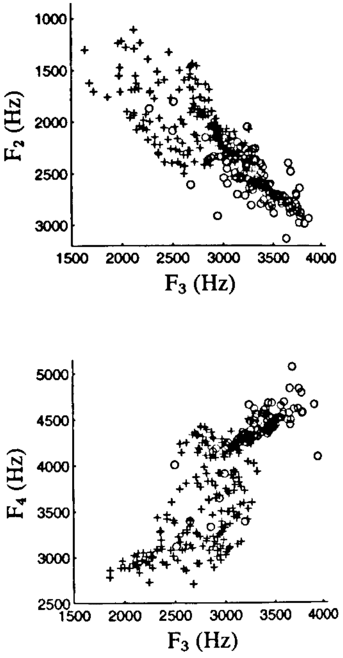

Figure 2: "Formant" values for retroflex constricitons with the tongue tip (+) and advanced tongue dorsum (○).

Modelling targets as probabilistic models offers two advantages. First, it is possible to estimated the likelihood of each clouds being produced by the each model $\mathcal{M}_j$. This gives some measure of confusion between acoustical results of different place of constriction. Let us call y_j the points the in cloud related to the model $\mathcal{M}_j$. The log likelihood of having y_k eeing produced by $\mathcal{M}_k$ is

$$\mathcal{L}(y_j|\mathcal{M}_k) = \sum_j \log[P(y_j|\mathcal{M}_k)]. \tag{2}$$

The greater $\mathcal{L}(y_j|\mathcal{M}_k)$, the more will have confusion at the acoustical space between the related places of constriction. We compute those values for 11 clouds, three for the tongue tip articulation and 8 for the tongue dorsum. The tongue dorsum can constrict as far as 5 cm back from the incisors, which means the soft palate region. The results are summarized in Fig. 3, in which the values for the likelihoods are shown as gray levels. We interpolated the data in order to improve the presentation. Darker regions correspond to high likelihoods. It is possible to see that some regions of confusions emerge from our data: between positions d_4 and d_7 (which corresponds to the hard palate), regions t_1 and t_2 (dental and alveolar) and d_2 and d_3 (advanced tongue dorsum). Velar regions and tongue tip retroflex configurations are quite separable from the others.

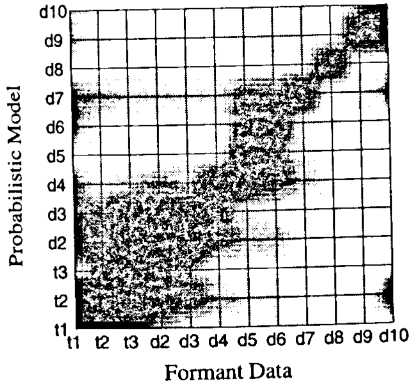

Figure 3: Log likelihhod of production of each cloud by each probabilistic model related to a place of constriction d_i stands for tongue dorsum and t_i for tongue dorsum. Darker regions mean high likelihood. Spacing of constrictions is 0.5 cm.

The second benefit of the probabilistic modelling is related to the way we can compute the error vector for the controller (Fig. 1). Indeed, for a given point in formant space y, the error vector is given simply by the gradient of the log likelihood with respect to y:

$$E = \frac{1}{P(y|\mathcal{M}_j)} \sum_{ij} g_{ij}|C_{ij}|^{-1/2} e^{-(y-y_{ij})^T C_{ij}^{-1}(y-y_{ij})}[C_{ij}^{-1}(y-y_{ij})]. \tag{3}$$

The error vectors generate a force field in the formant space which is transformed into changes in articulatory positions by the controller.

CONCLUSIONS

In this paper we showed how to obtain targets in the acoustical space for the stop components in the context of a model of motor control. We concentrated on describing the model and on how a probabilistic approach to the description of vowel–stop–vowel sequences can be useful. We showed that the description of clouds by mixtures of Gaussians yields interesting results, mainly related to the separability of the target regions for the different stop consonants produced by contact of the tongue to the hard- and soft-palate. Extensive simulations are planned for assessing the whole model.

ACKNOWLEDGEMENTS

We want to thank Jean-Luc SCHWARTZ for the insightful discussions and Louis-Jean BOË for making available the cineradiographic database. This research was supported by the European Community ESPRIT/BR grant # 6975 SPEECH MAPS and by GALILEO France-Italy project # 94290. The second author is recipient of a scholarship from CONACYT/Mexico.

REFERENCES

Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation* 6(2), 181–214.

Jordan, M. I. and D. E. Rumelhart (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science 16*, 307–354.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *J. Acous. Soc. Am. 73*(1), 322–335.

Klein, C. A. and C. H. Huang (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Man, Machines, and Cybernetics SMC-13*, 245–250.

Laboissière, R. (1993). Inversion and control of an articulatory model of the vocal tract: Recovering articulatory gestures from sounds. *J. Acous. Soc. Am. 93*(4), 2295.

Maeda, S. (1988). Improved articulatory model. *J. Acous. Soc. Am. 81*(S1), S146.

Stevens, K. N. and S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *J. Acous. Soc. Am. 64*(5), 1358–1368.

Sussman, H. M., H. A. McCaffrey, and S. A. Matthews (1991). An investigation of locus equations as a sources of relational invariance for stop place of categorization. *J. Acous. Soc. Am. 90*(3), 1309–1325.